Christoph Gröger[1]

Industrial analytics – An overview

Abstract: The digital transformation generates huge amounts of heterogeneous data across the industrial value chain, from simulation data in engineering, over sensor data in manufacturing to telemetry data on product use. Extracting insights from these data constitutes a critical success factor for industrial enterprises, e.g., to optimize processes and enhance product features. This is referred to as industrial analytics, i.e., data analytics for industrial value creation. Industrial analytics is an interdisciplinary subject area between data science and industrial engineering and is at the core of Industry 4.0. Yet, existing literature on industrial analytics is fragmented and specialized. To address this issue, this paper presents a holistic overview of the field of industrial analytics integrating both current research as well as industry experiences on real-world industrial analytics projects. We define key terms, describe typical use cases and discuss characteristics of industrial analytics. Moreover, we present a conceptual framework for industrial analytics that structures essential elements, e.g., data platforms and data roles. Finally, we conclude and highlight future research directions.

ACM CCS: Applied computing → Enterprise computing

Keywords: data analytics, machine learning, data management, data lake, industry, manufacturing

[1] **Corresponding Author: Christoph Gröger**, Robert Bosch GmbH, email: {prename.lastname}@de.bosch.com

1. Introduction

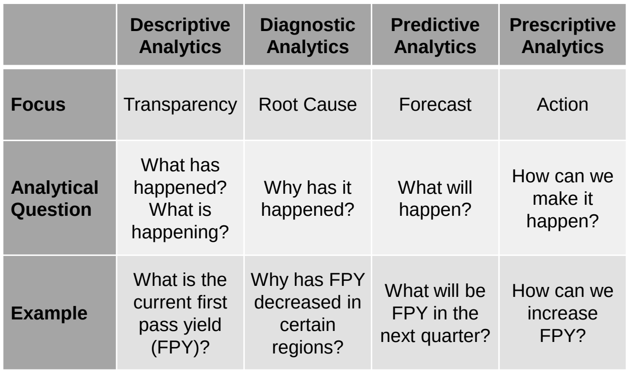
The global manufacturing industry is undergoing a fundamental change: from the production of physical goods to the delivery of data-driven processes and products. The digital transformation generates huge amounts of heterogenous data across the industrial value chain, from simulation data in engineering, over sensor data in manufacturing to telemetry data and social media postings on product use. Extracting insights and knowledge from these data constitutes a critical success factor for industrial enterprises, e.g., to optimize processes and enhance product features [1]. This is referred to as industrial analytics, i.e., data analytics for industrial value creation. Industrial analytics is an interdisciplinary subject area between data science and industrial engineering and is at the core of Industry 4.0 and the Industrial Internet of Things (IIoT) [2].

However, existing literature on industrial analytics is fragmented and specialized: for instance, there are works from the data science domain focusing on machine learning in manufacturing [1], [3], works from the computer science domain on data management for industrial enterprises [2], [4] and works from the industrial engineering domain on conceptual frameworks for industrial analytics and Industry 4.0 [5], [6]. Hence, an integrated view is lacking. To address this issue, this paper presents a holistic overview of the field of industrial analytics integrating both current research and industry experiences on real-world industrial analytics projects we published in [7]–[12]. As a basis, we define key terms on industrial analytics and industrial value creation in Section 2. Next, we describe typical use cases from industry practice in Section 3 and discuss characteristics of industrial analytics in comparison to web analytics in Section 4. Moreover, we present a conceptual framework for industrial analytics that structures essential elements, e.g., data platforms and data roles for industrial analytics, in Section 5. Finally, we conclude in Section 6 and highlight future research directions.

2. Industrial Analytics and Industrial Value Creation

The term *industrial analytics* generally refers to data analytics for industrial value creation with *Industry 4.0 analytics* [8] and *industrial intelligence* [13] being used as synonyms.

Data analytics comprises “theories, technologies, tools, and processes that enable an in-depth understanding and discovery of actionable insight into data” [14, p. 4]. That is, data analytics includes all kinds of data-driven analysis techniques, from reporting and online analytical processing (OLAP) over data mining and machine learning to stream analytics and complex event processing (see [14], [15] for details on these techniques). Thus, it also covers data-driven artificial intelligence, especially deep learning. All analysis techniques beyond reporting and OLAP are subsumed under the umbrella term advanced analytics [16]. Independent from the techniques used, there are four types of data analytics: descriptive analytics for transparency, diagnostic analytics for root causes, predictive analytics for forecasting as well as prescriptive analytics for action recommendations [17] (see Table 1).



	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Focus	Transparency	Root Cause	Forecast	Action
Analytical Question	What has happened? What is happening?	Why has it happened?	What will happen?	How can we make it happen?
Example	What is the current first pass yield (FPY)?	Why has FPY decreased in certain regions?	What will be FPY in the next quarter?	How can we increase FPY?

Table 1: Data analytics types

Industrial value creation constitutes the application domain of industrial analytics and is typically structured according to a process-centric or product-centric view. In the process-centric view, the industrial value chain [18] defines primary and secondary activities that are required for producing goods, e.g., logistics, marketing, service or procurement. In a product-centric view, which is used in the following, value creation is defined by the industrial product lifecycle [19]. It comprises all phases from product development over sales and distribution to retirement and recycling of a product (see left part of Figure 1).

Industrial analytics is at the core of *Industry 4.0* [20], respectively the IIoT [21]. These approaches generally refer to the next generation of industrial value creation based on the comprehensive use of IoT technology and cyber-physical systems, e.g., by flexible decentral execution of manufacturing processes. The overall goal is to realize a complete digital interconnection of all processes and objects in industrial value creation, within the enterprise and across the entire supply chain. The resulting data are manifold and comprise internal and external data, batch and stream data as well as structured and unstructured data on all aspects of industrial processes and products. Examples include sensor data from machines in manufacturing, quality test data from product development or customer sentiments from web blogs.

The *goal of industrial analytics* is to exploit the business value of these data by realizing data-driven products and services as well as data-driven processes.

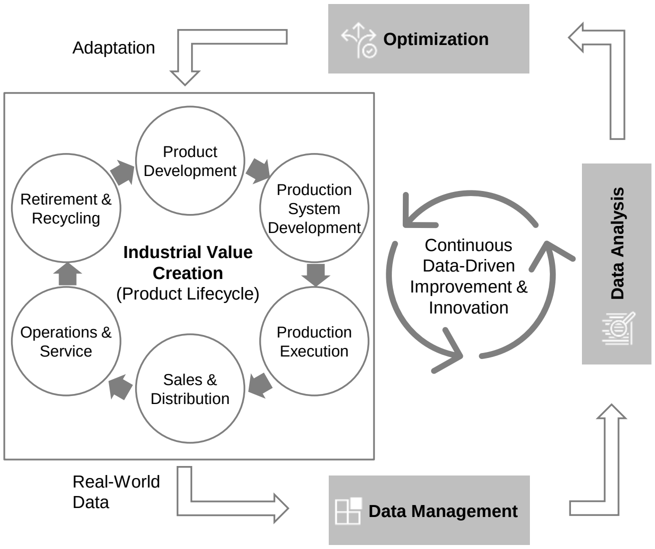


Figure 1: Industrial analytics cycle with three phases

Data-driven products and services comprise, e.g., enhanced products and value-added services such as self-optimizing machines and connected car services. Data-driven processes refer to data-driven transparency, root cause analysis, prediction and optimization of processes, e.g., with predictive maintenance or prescriptive process analytics.

To achieve this goal, industrial analytics can be represented as a learning cycle with three phases (see [8] for elements of the cycle). As shown in Figure 1, real-world data, including metadata, from the product lifecycle constitute the starting point. Then, *data management* focuses on the acquisition, integration and preprocessing of these data to establish the data basis for further analyses, e.g., by storing and cleansing data in data lakes. *Data analysis* refers to the evaluation of data by applying various analysis techniques, e.g., reporting and machine learning. *Optimization* relates to deriving and implementing specific improvement actions based on the generated analysis results to adapt processes or products, e.g., adapting product features based on the analysis of product usage data. In the end, the adapted process or product is executed or used again and generates new data as a foundation for the next iteration of the cycle. Thereby, it enables continuous data-driven improvement and innovation.

3. Typical Use Cases

There is a wide range of industrial analytics use cases across the value chain and across the product lifecycle. On the one hand, there are classical use cases focusing on descriptive analytics, especially reporting, across all areas of an industrial enterprise. This includes, for instance, key performance indicator (KPI) reporting in manufacturing, human resources or sales mainly based on enterprise resource planning (ERP) data and is extensively covered in various works on business intelligence [22]. On the other hand, there are advanced use cases that exploit additional data generated by digitalization efforts and that make use of advanced analytics, especially data mining and machine learning. In the following, we give an overview of three advanced industrial analytics use cases that are typical in industry practice (see Table 2).

The goal of *predictive machine maintenance* [23] is to optimize activities and schedules of machine maintenance on the shop floor using historic machine data and maintenance data, e.g., from a manufacturing execution system (MES). This goal is twofold: it is about avoiding unplanned machine downtimes due to machine failures as well as about avoiding unnecessary maintenance on quality related parts, e.g., molding tools. There are various implementation approaches depending on the individual case. In principle, data mining and machine learning methods are used to generate a prediction model that forecasts remaining operating time of the machine or the likelihood of failures in defined time periods considering performed maintenance activities and the machine’s state [24]. On this basis, maintenance cycles can be optimized. Challenges in industry practice mainly result from missing availability and quality of data on maintenance activities. Depending on the individual machine, maintenance activities may be recorded manually in spreadsheets only or even not at all. In addition, sufficient historic data on machine failures is required to develop a valid prediction model. Typically, these data are imbalanced with little number of failures and thus make prediction model development challenging (see [1] for imbalanced data in manufacturing).

The aim of *predictive process quality* [25] is to reduce scrap rates and enhance first-pass-yield of a manufacturing process. The process can be comprised of several manual or automated process steps involving various resources, especially machines and tools. At this, there can be various complex interdependencies between process parameters and machine parameters as well as between tool characteristics and material characteristics across the entire process. These interdependencies influence process quality, which is typically measured by end-of-line tests of produced parts. Thus, predictive process quality implementations are based on a wide range of source data, especially material data on input materials, quality data on test results as well as machine and process data, e.g., from MESs and machine sensors. On this basis, data mining and machine learning methods are used for developing prediction models that enable forecasting the likely end-of-line quality of parts in process as well as identifying root causes for insufficient quality across the entire process. Implemented online during process execution, parts in process can then be

	Predictive Machine Maintenance	Predictive Process Quality	Engineering in the Loop
Goal	Optimizing machine maintenance	Reducing scrap in manufacturing	Improving product design based on real-world product usage
Object of Analysis	Machine	Process	Product
Product Lifecycle Phases	Production system development, production execution	Production system development, production execution	Product development, operations & service
Source Data	Maintenance data, machine data	Material data, quality data, process data, machine data	Engineering simulation data, master data, product usage data
Analytics Types	Predictive	Diagnostic, predictive	Descriptive
Techniques	Data mining & machine learning	Data mining & machine learning	Reporting & OLAP, exploration
Challenges in Practice	Data availability, data quality, imbalanced data	Data integration, imbalanced data	Data availability, data quality

Table 2: Typical industrial analytics use cases

sorted out or reworked before they cause additional quality costs and root causes for decreasing quality can be identified. Challenges in industry practice particularly refer to data integration and the handling of imbalanced data. Linking quality results of a specific part to the corresponding process and machine parameters, requires references to individual parts in the data and makes data integration complex and costly. Moreover, in case there is only a small number of produced parts with insufficient quality, imbalanced data result and make prediction model development complex.

Engineering in the loop aims at improving product design based on real-world product usage and represents a use case of data-driven engineering [26]. In principle, data on real-world product usage, e.g., collected by IoT-enabled devices, is analyzed together with engineering data closing the loop between product development and operations in the product lifecycle. By comparing product design, i.e., designated usage, and real-world usage, product characteristics can be further developed for the next product generation. For instance, in case real-world usage data reveals that product use differs with respect to the geographical location of the customer, product configurations can be adapted accordingly. The basic variant of engineering in the loop is descriptive and makes use of exploration as well as reporting and OLAP techniques for data analysis. Advanced variants employ data mining and machine learning for identifying typical usage patterns for certain products. Challenges in practice mainly refer to data availability, i.e., collecting and accessing real-world product usage data from customers, as well as quality of these data. The latter is typically difficult to assess and varies widely as do the conditions of product use do. For instance, some products may only be used for demonstrating or testing purposes at certain customers, others may be used productively.

4. Characteristics

Many concepts, architectures and technologies for industrial analytics originate from the field of web analytics [2]: for instance, large-scale data pipelining in data lakes with Apache Hadoop[1] and Apache Spark[2]. However, such approaches cannot simply be adopted one-to-one but have to be revisited and customized according to the characteristics of the field of industrial analytics. These characteristics shown in Table 3 are detailed in the following in a qualitative comparison to web analytics (see [27] for web analytics characteristics).

The *goal* of web analytics is to analyze web user behavior for user experience improvement and digital marketing, e.g., with individual product recommendations in web shops or target-group-oriented ads. Accordingly, internet companies such as Amazon, Facebook and Google are major *business players* in the field of web analytics. In contrast, industrial analytics aims at analyzing and optimizing industrial products, processes and services. Industrial enterprises such as Bosch, Airbus or Siemens are central players in this field.

Main data sources of web analytics are web server logs, page tagging data as well as data of the respective web application, e.g., the web shop with customer orders and products. In addition, external data, e.g., socio-demographic data, are used to enrich the data. Regarding industrial analytics, enterprise IT systems such as ERP systems, MESs as well as customer relationship management (CRM) systems are central data sources (see [28] for details on enterprise IT systems). In addition, there are IoT data, e.g., from sensors in manufacturing

[1] https://hadoop.apache.org

[2] https://spark.apache.org

	Web Analytics	**Industrial Analytics**
Goal	Analyzing web user behavior for user experience improvement and digital marketing	Analyzing and optimizing industrial products, services and processes
Business Players	Internet companies	Industrial enterprises
Main Data Sources	Web server logs, page tagging data, web application data, external data (socio-demographic, …)	Enterprise IT systems (ERP, MES, CRM, …), IoT data, external data (supplier data, …)
Entity Identification	Comparatively simple (e.g., using fingerprinting or cookies)	Comparatively complex (e.g., identification of individual parts in a manufacturing process)
Data Heterogeneity	Comparatively low (e.g., standardized web server log formats)	Comparatively high (e.g., individual MES data models per production line)
Data Integration	Comparatively simple	Comparatively complex
Reliability Requirements	Comparatively low (e.g., single event in a click stream may be neglected)	Comparatively high (e.g., robust processing of safety-related events in manufacturing)

Table 3: Comparison of web analytics and industrial analytics

machines and products, as well as external data, e.g., supplier data.

Entity identification refers to determining and tracking the individual objects of interest, also across different data sources. In the case of web analytics, determination and tracking of individual web users is comparatively simple due to established techniques such as cookies and browser fingerprinting. By contrast, entity identification can be far more complex in industrial analytics due to the complexity and heterogeneity of products and processes in industrial value creation. For instance, tracking individual metal parts with imprinted identification codes can be challenging when machining steps such as grinding soil the codes. Depending on the type of manufacturing process, identification of individual parts can be even impossible or economically not feasible, e.g., in the case of large-scale batch production.

Data heterogeneity is considered comparatively low in web analytics due to wide-spread data standards for major data sources, e.g., standardized web server log formats such as NCSA[1]. In industrial analytics, data heterogeneity is considered high due to the wide variety of data sources and their adaptation to the heterogeneity of industrial processes, especially in manufacturing. For instance, ERP systems and MESs are typically tailored to the specific type of production, e.g., batch production or single-piece production. Moreover, in complex manufacturing processes, data models of individual MESs are customized to the specific production line that they manage.

Data integration is considered rather simple in web analytics compared to industrial analytics due to simpler entity identification and lower data heterogeneity. Based on identified web users and standardized source data, integration of various user-related data, e.g., web application data and click streams for determined users, are facilitated. Data integration in industrial analytics is more complex because many data sources lack data standards and entity references. For instance, measured machine vibrations can be an indicator for manufacturing imprecisions and maintenance needs. Yet, linking vibration sensor data with the respective parts manufactured is challenging as the sensor data misses direct references to the parts produced at certain vibration levels. Thus, integration has to be done manually by using timestamps, for example.

Reliability requirements for data and analytics pipelines in web analytics are considered comparatively low in contrast to industrial analytics. For instance, single events in a click stream may be lost due to a failure of the processing pipeline without significantly impacting business. By contrast, neglecting single events from a manufacturing machine can be economically relevant when these data refer to safety-relevant events and have to be documented for compliance reasons. Hence, corresponding data pipelines have to be highly available and fault tolerant.

All in all, this qualitative comparison shows that industrial analytics differs from web analytics in key aspects. Thus, it constitutes a separate application domain of data analytics and requires the development of tailored concepts, architectures and platforms, e.g., for the integration of domain knowledge in manufacturing data pipelines [29].

5. Industrial Analytics Framework

There are two frameworks for industrial analytics and Industry 4.0 that are widespread in practice: the Reference Architectural Model Industrie 4.0 (RAMI4.0) and the IIoT Reference Architecture with its analytics framework.

[1] https://ibm.co/3Hoi3Kd

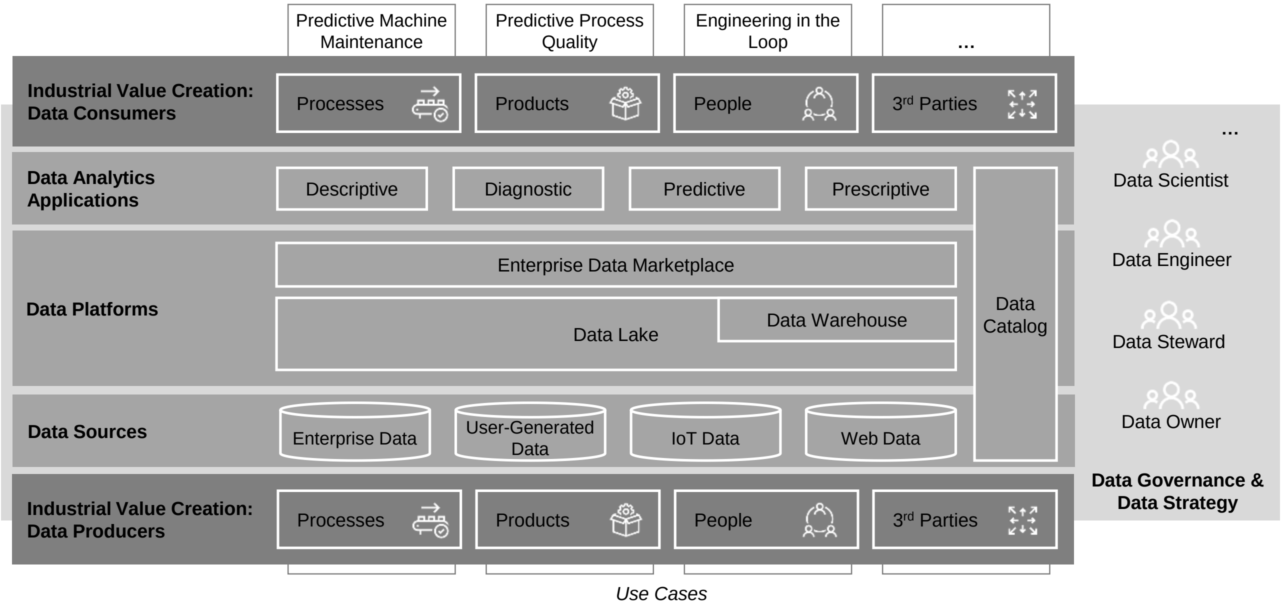


Figure 2: Industrial analytics framework (extended from [11])

The *RAMI4.0 framework* [5] takes a conceptual view and integrates three dimensions on data abstraction, manufacturing hierarchy levels and the lifecycle of objects. It particularly focuses on digital twins, data models and value chain integration without detailing IT implementation aspects. The *IIoT Reference Architecture* [6] focuses on the integration of industrial control and shop floor systems with business processes and enterprise IT systems. The included analytics framework presents some generic requirements and implementation technologies for industrial analytics, e.g., cloud and edge computing or complex event processing.

Our industrial analytics framework complements these works by combining conceptual and implementation-related aspects. Moreover, it takes a data-centric view on industrial analytics because data challenges constitute the major obstacle to leverage data analytics in industrial value creation. The framework constitutes a refined version of the data ecosystem for industrial enterprises that we developed in our previous work [11].

As shown in Figure 2, the framework combines IT-technical and organizational aspects, e.g., data analytics applications and data governance, with various kinds of data platforms at its core. It is based on the concept of a self-organizing and loosely coupled ecosystem for the sharing of data between all kinds of data producers and data consumers in industrial enterprises (see [11] for the elements presented in the following sections).

5.1. Data Producers and Data Consumers

Data producers and data consumers are resources or actors that generate or use data. There are four major kinds of data producers respectively consumers in industrial enterprises:

- *Processes* comprise all kinds of industrial processes including underlying resources, e.g., manufacturing processes including machines, sales processes or engineering processes.
- *Products* refer to manufactured goods such as injectors, drives or electrical appliances.
- *People* comprise human actors, especially employees and customers of the enterprise
- *Third parties* refer to all kinds of actors and resources outside of the organizational scope of the enterprise, e.g., suppliers and payment providers.

5.2. Data Sources

Data sources represent the technical kind and the sources of data generated by data producers. There are four kinds of data sources in industrial enterprises:

- *Enterprise data* comprise all data generated by enterprise IT systems across the industrial product lifecycle, e.g., ERP systems, MESs, product lifecycle management (PLM) systems or computer-aided design (CAD) systems.
- *User-generated data* refer to data directly produced by human actors, e.g., documents, spreadsheets or social media postings.
- *IoT data* comprise data from sensor-equipped physical objects, e.g., manufacturing machine data or sensor data from IoT-enabled products.
- *Web data* refer to all data from the web, except user-generated data, e.g., open life science data or open government data.

All kinds of data comprise structured, semi-structured and unstructured data as well as batch and stream data.

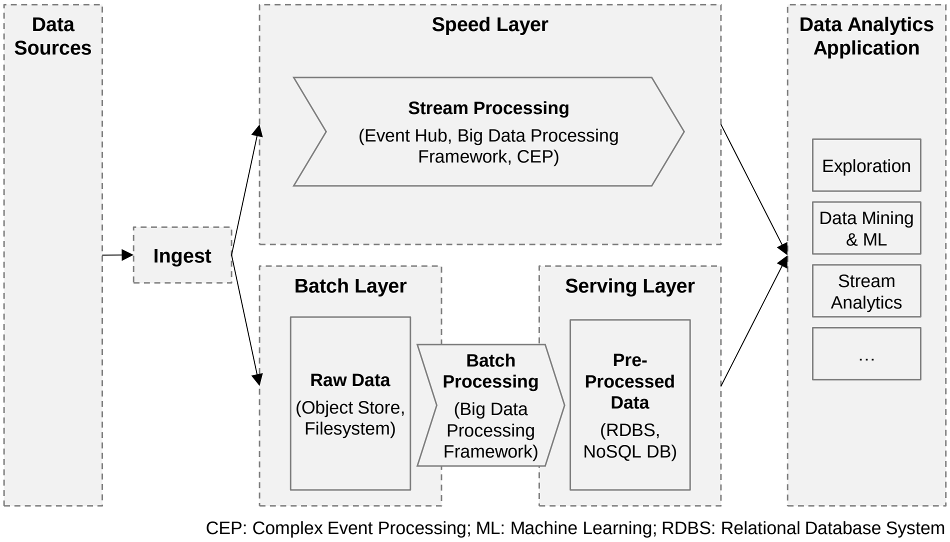

Data Sources
Ingest
Speed Layer
Stream Processing
(Event Hub, Big Data Processing Framework, CEP)
Batch Layer
Raw Data
(Object Store, Filesystem)
Batch Processing
(Big Data Processing Framework)
Serving Layer
Pre-Processed Data
(RDBS, NoSQL DB)
Data Analytics Application
Exploration
Data Mining & ML
Stream Analytics
…
CEP: Complex Event Processing; ML: Machine Learning; RDBS: Relational Database System

Figure 3: Lambda architecture for data lakes with typical technologies

5.3. Data Platforms

Data platforms constitute the core of the framework and represent the technical basis for data storage and processing from all kinds of data sources with the aim to make data available for all kinds of data analytics applications. That is, data platforms realize the phase of data management in the industrial analytics cycle.

There are four types of data platforms: apart from the data warehouse as established platform for the analysis of structured data, the need for advanced analytics on both structured and unstructured data has revealed the data lake. Moreover, the necessity for managing metadata and for democratizing existing data in data lakes and data warehouses has given birth to the data catalog and the enterprise data marketplace as most recent platform developments in industry practice. In the following, we give a short overview of these platforms and point out essential architecture aspects and realization technologies.

Data warehouses [30] represent the basis for classical business intelligence, i.e., reporting and OLAP on aggregated structured enterprise data such as ERP data. The goal is to enable consistent enterprise-wide KPI reporting at all management levels. Architectures and technologies are mature and proven using various data modelling approaches, e.g., multidimensional modelling, implemented in relational database systems mostly on-premises.

Data lakes [31] emerged as a complement to data warehouses. They constitute scalable platforms for storing and processing all kinds of data – structured, semi-structured and unstructured – in their raw form for flexible analysis, especially advanced analytics such as machine learning and stream analytics.

Architectures and technologies of data lakes are mostly driven by experimental developments in practice and significant efforts of the open-source community in big data processing. Typically, data lake architectures are based on a polyglot approach combining batch and stream processing as well as relational and non-relational database systems according to the so-called Lambda architecture [32] with batch, speed and serving layer (see Figure 3). Technologies are manifold comprising distributed filesystems and objects stores for large-scale raw data storage, big data processing frameworks such as Apache Spark[1] for scalable data processing and a variety of relational and NoSQL database systems for use-case-specific data provisioning in the serving layer, e.g., MySQL[2] or Apache Cassandra[3]. Deployments of data lakes are typically done both on-premises and in the cloud to leverage scalability and pay-per-use models for peak workloads in public clouds.

As data lakes are increasingly being used for reporting use cases and as they make use of similar source data as data warehouses, a current trend in industry practice is to combine data lakes and data warehouses in integrated data platforms. The data warehouse then becomes part of the serving layer of the data lake and gets feed with raw data provided by the batch layer.

Data catalogs [33] are metadata-based inventories of data to foster data transparency across operational and analytical IT systems. They store and integrate different kinds of metadata from enterprise IT systems, data lakes

[1] https://spark.apache.org
[2] https://www.mysql.com
[3] https://cassandra.apache.org

and data warehouses as well as from data analytics applications. For instance, technical metadata such as names and types of columns in relational database systems are captured just like business metadata such as the business meaning of columns or the responsible data owner of a certain table. In addition, operational metadata are stored comprising details on data transformations and accesses, e.g., regarding extraction, transformation and loading (ETL) jobs.

On this metadata basis, data catalogs support discovery and understanding of data as well as impact and lineage analyses, e.g., to determine the provenance of datasets in a data lake or investigate the impact of changes in source systems on related ETL processes.

Regarding architectures and technologies, data catalogs are typically provided as standard software based on a classical three-layer architecture. The data layer comprises a metadata repository using a relational or NoSQL database system, the application layer provides the application logic for capturing, storing and integrating metadata and the presentation layer is typically realized as a web application to facilitate simple user access.

There are commercial and open-source data catalog products, e.g., Informatica Enterprise Data Catalog[1], Collibra Data Catalog[2] or Amundsen Data Catalog[3]. They typically make use of a custom data model for structuring metadata. A current trend is to exploit semantic and graph-based technologies such as the resource description framework and triple stores to manage metadata and foster the standardization of metadata models [34].

Enterprise data marketplaces [11] are a novel type of data platform currently emerging in industry practice. They represent metadata-based self-service platforms that connect data producers and data consumers within an enterprise. The goal is to foster data democratization, i.e., making data available to everyone in an organization – whilst ensuring compliance. To this end, they heavily make use of metadata and thus extend the metadata provided by the data catalog. For data producers, enterprise data marketplaces provide self-services for data provisioning and curation, especially in data lakes. For instance, when a manufacturing engineer wants to provide sensor data from a recently installed machine sensor for other data scientists without involving the IT department for an ETL project. For data consumers, enterprise data marketplaces provide self-services for data discovery and data preparation whilst ensuring compliance, e.g., by realizing automated workflows for compliant access of personal-related data.

Architectures and technologies for enterprise data marketplaces are subject to current research [11]. Semantic technologies are a promising candidate for managing metadata and services in enterprise data marketplaces to enable flexible extension and interoperability of services and (meta)data. Besides, fusing data catalogs and enterprise marketplaces into integrated metadata platforms is a valuable direction of future work.

5.4. Data Analytics Applications

Data analytics applications make use of data provided by data platforms, especially by data lakes and data warehouses. There are four kinds of data analytics applications depending on the type of analytics: descriptive, diagnostic, predictive and prescriptive applications (see Section 2).

Data analytics applications realize defined use cases, e.g., engineering in the loop, for defined data consumers, e.g., product engineers. For this purpose, they make use of various data-driven analysis techniques, especially reporting and OLAP, exploration as well as data mining and machine learning as explained in the following.

Reporting and OLAP [22] refer to the multidimensional analysis of metrics, i.e., facts, according to defined dimensions, e.g., analyzing scrap rates for different kinds of products, production lines and factories. Resulting applications are typically descriptive or diagnostic and are implemented using reporting and OLAP platforms, e.g., SAP Business Objects[4] or IBM Cognos[5].

Data exploration [14] refers to the interactive investigation of data. It is either done in code-free manner using self-service visualization tools such as Tableau or code-based using interactive notebooks such as Jupyter[6]. Generated results can then be distributed using dashboards or reports.

Data mining and machine learning [15] generally refer to the (semi-)automatic discovery of valuable patterns from data. They comprise descriptive and diagnostic methods, e.g., for clustering and association rule discovery, as well as predictive methods for classification and regression. There is a huge variety of algorithms to implement these methods including algorithms for artificial neural networks and deep learning. Accordingly, tools for data mining and machine learning vary greatly: from code-free visual tools with a pre-defined set of methods and algorithms such as Knime[7] to code-based environments with a comprehensive library ecosystem such as Python[8]. At this, managing the entire

[1] https://www.informatica.com/products/data-catalog.html
[2] https://www.collibra.com/data-catalog
[3] https://www.amundsen.io
[4] https://www.sap.com/products/bi-platform.html
[5] https://www.ibm.com/products/cognos-analytics
[6] https://jupyter.org
[7] https://www.knime.com
[8] https://www.python.org/

development process and deploying generated models in operational systems is essential, e.g., when using a prediction model for supplier risk assessment in an ERP system. For this issue, so-called machine learning operations (MLOps) platforms such as MLflow[1] come into play.

For prescriptive applications, there are no dedicated analysis techniques. Typically, various techniques such as data mining and machine learning as well as metric calculations are combined according to the specific use case, e.g., prescriptive process analytics [35].

5.5. Data Governance and Data Strategy

Data governance refers to organizational roles, rights and responsibilities to treat data as an enterprise asset [36]. It is specifically about data owners and data stewards as well as data analytics roles such as data scientists and data engineers as briefly described in the following.

Data owners [36] are responsible for the security, compliance and quality for a certain kind of data, e.g., MES data, from a business point of view. *Data stewards* [36] are tasked by data owners to define and implement corresponding policies and technical procedures, e.g., to implement reference data models for manufacturing data and ensure defined data quality levels.

Data engineers and *data scientists* [37] are roles in the context of data analytics projects. In short, data engineers focus on preparing data by developing data pipelines in data lakes and data warehouses. On this basis, data scientists focus on the actual analysis, e.g., by applying various machine learning algorithms. It is to be noted that the roles of data stewards and data engineers partially overlap and further work is needed to define consistent models for data ownership and data stewardship in industrial enterprises as described in our previous work [11].

Beyond data governance, a *data strategy* [38] is about how to generate busines value from data. That is, a data strategy defines both key conditions for data management and data governance as well as concepts for data-driven business, e.g., novel business models. Data strategies are currently developed in both industrial and digital companies and are an object of current information systems research (see [38] for more details on data strategies).

6. Conclusion and Future Directions

Industrial analytics is an interdisciplinary subject area between data science and industrial engineering. The discussion of essential characteristics in comparison to web analytics reveals that industrial analytics constitutes a separate application domain of data analytics requiring the development of tailored concepts, architectures and platforms. Data challenges represent the major obstacle to leverage industrial analytics in practice. Therefore, we designed an industrial analytics framework that specifically focuses on different kinds of data platforms. Valuable directions of future work comprise the development of integrated platforms for data management combining data lakes and data warehouses as well as of integrated platforms for metadata management fusing data catalogs and enterprise data marketplaces. Moreover, designing organizational concepts for data democratization, e.g., with incentives for data producers to share their data freely within the enterprise, are a crucial success factor.

Literature

[1] A. Dogan and D. Birant, “Machine learning and data mining in manufacturing,” *Expert Systems with Applications*, vol. 166, pp. 1–22, 2021.

[2] Y. Cui, S. Kara, and K. C. Chan, “Manufacturing big data ecosystem: a systematic literature review,” *Robotics and Computer Integrated Manufacturing*, vol. 62, p. Article 101861, 2020.

[3] M. Sharp, R. Ak, and T. Jr. Hedberg, “A survey of the advancing use and development of machine learning in smart manufacturing,” *Journal of Manufacturing Systems*, vol. 48, pp. 170–179, 2018.

[4] A. Ismail, H.-L. Truong, and W. Kastner, “Manufacturing process data analysis pipelines: a requirements analysis and survey,” *Journal of Big Data*, vol. 6, no. 1, pp. 1–26, 2019.

[5] Plattform Industrie 4.0, “RAMI4.0 – a reference framework for digitalisation.” 2018. [Online]. Available: https://www.plattform-i40.de/IP/Redaktion/EN/Downloads/Publikation/rami40-an-introduction.pdf?__blob=publicationFile&v=7

[6] N. Anderson *et al.*, “The Industrial Internet of Things Volume T3: Analytics Framework.” 2017. [Online]. Available: https://www.iiconsortium.org/pdf/IIC_Industrial_Analytics_Framework_Oct_2017.pdf

[7] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang, “Modeling data lakes with data vault: practical experiences, assessment, and lessons learned,” in *Proceedings of the International Conference on Conceptual Modeling (ER)*, Cham, 2019, pp. 63–77.

[8] C. Gröger, “Building an industry 4.0 analytics platform,” *Datenbank-Spektrum*, vol. 18, no. 1, pp. 5–14, 2018.

[9] E. Wagner, B. Keller, P. Reimann, C. Gröger, and D. Spath, “Advanced Analytics for Evaluating Critical Joining Technologies in Automotive Body Structures and Body Shops,” in *Proceedings of the CIRP Conference on Intelligent Computation in Manufacturing Engineering (CIRP ICME)*, 2021, p. to appear.

[10] A. Birk, Y. Wilhelm, S. Dreher, C. Flack, P. Reimann, and C. Gröger, “A Real-World Application of Process Mining for Data-Driven Analysis of Multi-Level Interlinked

[1] https://mlflow.org

Manufacturing Processes," *Procedia CIRP*, vol. 104, pp. 417–422, 2021.

[11] C. Gröger, "There Is No AI Without Data," *Communications of the ACM*, vol. 64, no. 11, pp. 98–108, 2021.

[12] L. Kassner *et al.*, "The Stuttgart IT Architecture for Manufacturing. An Architecture for the Data-Driven Factory," in *Enterprise Information Systems (ICEIS) 2016. Revised Selected Papers*, S. Hammoudi, L. Maciaszek, M. Missikoff, O. Camp, and J. Cordeiro, Eds. Cham: Springer, 2017, pp. 53–80.

[13] H. Lasi, "Industrial intelligence - a business intelligence-based approach to enhance manufacturing engineering in industrial companies," *Procedia CIRP*, vol. 12, pp. 384–389, 2013.

[14] L. Cao, "Data science: a comprehensive overview," *ACM Computing Surveys*, vol. 50, no. 3, pp. 1–42, 2017.

[15] J. Han, M. Kamber, and J. Pei, *Data mining: Concepts and techniques*, 3rd ed. Amsterdam: Morgan Kaufmann, 2012.

[16] R. Bose, "Advanced analytics: opportunities and challenges," *Industrial Management & Data Systems*, vol. 109, no. 2, pp. 155–172, 2009.

[17] L. Kart, A. Linden, and W. R. Schulte, "Extend your portfolio of analytics capabilities. Gartner research note G00254653." 2013.

[18] M. E. Porter, *Competitive Advantage*. New York: Free Press, 1985.

[19] J. Stark, *Product Lifecycle Management*, 4th ed. Cham: Springer, 2020.

[20] T. Bauernhansl, "Die Vierte Industrielle Revolution – Der Weg in ein wertschaffendes Produktionsparadigma," in *Industrie 4.0 in Produktion, Automatisierung und Logistik. Anwendung, Technologien, Migration*, T. Bauernhansl, M. t. Hompel, and B. Vogel-Heuser, Eds. Wiesbaden: Springer Vieweg, 2014, pp. 5–35.

[21] S. Jeschke, C. Brecher, T. Meisen, D. Özdemir, and T. Eschert, "Industrial internet of things and cyber manufacturing systems," in *Industrial internet of things*, S. Jeschke, C. Brecher, H. Song, and D. Rawat, Eds. Cham: Springer, 2017, pp. 3–19.

[22] H. Baars and H.-G. Kemper, *Business Intelligence & Analytics*, 4th ed. Wiesbaden: Springer Vieweg.

[23] H. M. Hashemian and W. C. Bean, "State-of-the-Art Predictive Maintenance Techniques," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 10, pp. 3480–3492, 2011.

[24] M. Paolanti, L. Romeo, A. Felicetti, A. Mancini, E. Frontoni, and J. Loncarski, "Machine learning approach for predictive maintenance in industry 4.0," in *Proceedings of the IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, 2018, pp. 1–6.

[25] C. Gröger, *Advanced Manufacturing Analytics - Datengetriebene Optimierung von Fertigungsprozessen*. Lohmar: Josef Eul, 2015.

[26] J. Trauer, S. Schweigert-Recksiek, L. O. Okamoto, K. Spreitzer, M. Mörtl, and M. Zimmermann, "Data-Driven Engineering – Definitions and Insights from an Industrial Case Study for a New Approach in Technical Product Development," in *Proceedings of NordDesign*, 2020, pp. 1–12.

[27] G. Zheng and S. Peltsverger, "Web Analytics Overview," *Encyclopedia of Information Science and Technology*. IGI Global, Hershey, pp. 7674–7683, 2015.

[28] K. C. Laudon and J. P. Laudon, *Management information systems. Managing the digital firm*, 15th ed. Harlow: Pearson Education, 2018.

[29] V. Hirsch, P. Reimann, and B. Mitschang, "Exploiting domain knowledge to address multi-class imbalance and a heterogeneous feature space in classification tasks for manufacturing data," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3258–3271, 2020, doi: 10.14778/3415478.3415549.

[30] W. H. Inmon, *Building the Data Warehouse*, 4th ed. Indianapolis: Wiley, 2005.

[31] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang, "Leveraging the data lake: current state and challenges," in *Proceedings of the International Conference on Big Data Analytics and Knowledge Discovery (DaWaK)*, Cham, 2019, pp. 179–188.

[32] N. Marz and J. Warren, *Big data: Principles and best practices of scalable real-time data systems*. Shelter Island: Manning, 2015.

[33] E. Zaidi, G. d. Simoni, R. Edjlali, and A. D. Duncan, "Data Catalogs Are the New Black in Data Management and Analytics. Gartner Research Note G00338777." 2017.

[34] H. Dibowski and S. Schmid, "Using Knowledge Graphs to Manage a Data Lake," in *Proceedings of the Annual Conference of the German Informatics Society (Informatik)*, Bonn, 2020, pp. 41–50.

[35] C. Gröger, H. Schwarz, and B. Mitschang, "Prescriptive Analytics for Recommendation-based Business Process Optimization," in *Proceedings of the 17th International Conference on Business Information Systems (BIS), Larnaca, Cyprus, 21-23 May, 2014*, 2014, pp. 25–37.

[36] R. Abraham, J. Schneider, and J. v. Brocke, "Data governance: a conceptual framework, structured review, and research agenda," *International Journal of Information Management*, vol. 49, pp. 424–438, 2019.

[37] S. Michalczyk, M. Nadj, A. Maedche, and C. Gröger, "Demystifying Job Roles in Data Science: A Text Mining Approach," in *Proceedings of the 29th European Conference on Information Systems (ECIS)*, 2021, pp. 1–17.

[38] I. Gür, M. Spiekermann, M. Arbter, and B. Otto, "Data Strategy Development: A Taxonomy for Data Strategy Tools and Methodologies in the Economy," in *Proceedings of Wirtschaftsinformatik (WI)*, 2021, pp. 1–15.